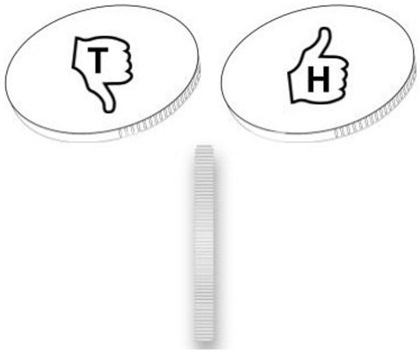
Statistics of "TRUE" and "FALSE"



In a game of cricket the team that wins the toss gets to choose to bat or to bowl. When a coin is tossed into the air, it lands showing one of the two faces namely the head or the tail. {Picture above left} If the coin is tossed a number of times, would the number of "heads" be equal to the number of "tails"? We will ignore the very rare chance that the coin actually stands on its edge. If they are equal, the coin is called "fair". But, how many tosses should we count? Everyone knows that we do not get two heads out of every four tosses. So a few tosses are not clearly sufficient. A further question. For the coin to be called fair, should the number of heads and tails be exactly equal or only approximately equal? Bias of a coin is the number of heads divided by the number of tails. The bias of a fair coin is 1. The number of heads and tails obtained after a large number of tosses would be equal. If the coin has a bias less than one there are more tails than heads. For example a coin with a bias of 0.33 will show two tails for every head. A coin with bias higher than 1 will show more heads. If the bias is 2, there will be two heads for every tail.

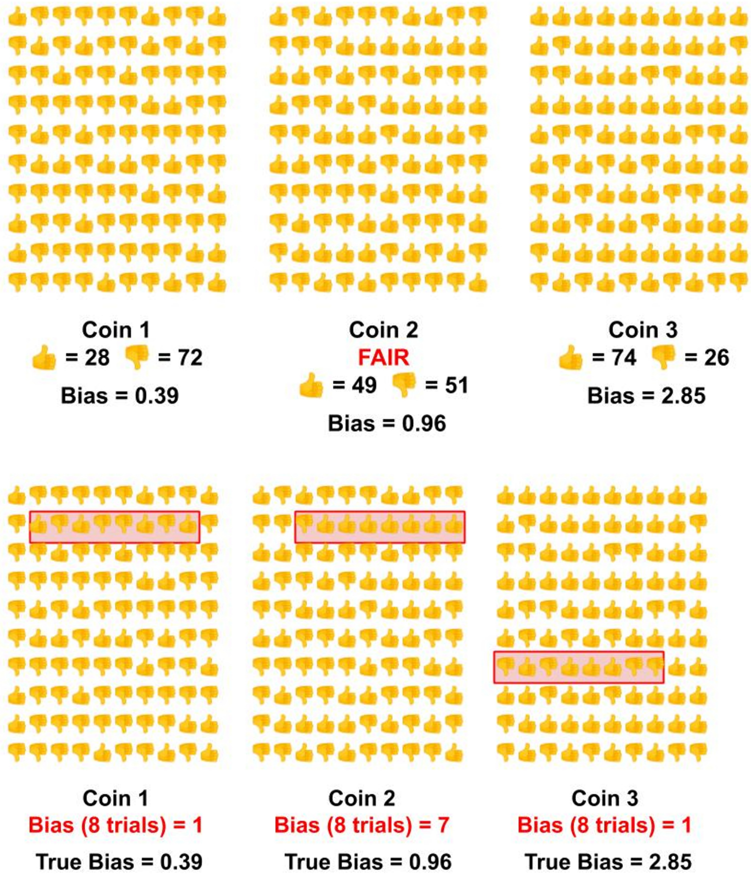
Coins with bias different from 1 may not actually exist. It is also very difficult to develop a skill to make the coin land head or tail as desired. But it is possible to prepare mathematical descriptions of biased coins and obtain the results of tossing on a computer. Such an examination is very useful for evaluating the strength of experimental results. In many experiments, the results are available not as a distribution with a mean, median and mode but as one of a pair; YES-NO; SUCCESS-FAILURE; RIGHT-WRONG etc. Often this is the result of simplification. For example the result of a blood test for a disease may state that

the patient IS suffering from the disease or that he or she is NOT suffering from the disease. Perhaps the biochemical reactions give a Gaussian result and the comparison between the two Gaussians is simplified as a positive or negative test result.

Any test or experimental result has a known probability of error. This is the most important aspect of a scientific result. In the example of the blood test, the scientist would tell even before the test that the results have an “x” percent chance of error. This admission of a chance of error and defining that chance is not to be seen in non scientific claims. The lack of a defined error can be an easy way to recognize unscientific claims.

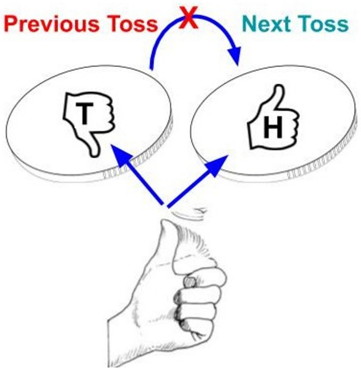
The error of the test is the standard deviation discussed earlier. When the results are in the form of yes and no, the uncertainty in results is related to the bias. That is the reason we are discussing the tossing of coins here. For example, consider a blood test that the doctor assures is 90% reliable. This is mathematically the same as a coin with a bias of 9. For every nine heads there is a tail. For every 9 successful diagnoses of the disease there is one false diagnosis. How did the scientists come to the conclusion that the test is 90% reliable? Through experimental trials. How many experiments have to be done? What number of coin tosses will result in what confidence in the value of the bias? Here, we cannot discuss such questions of advanced statistics in detail here. But some analysis of coin tosses will still be very useful.

Three coins with different bias values were "tossed" on a computer and the results presented. {Picture on next page} Heads are shown as a thumbs up emoticon and tails as a thumbs down emoticon. The results shown in the middle appear to be from a fair coin. The bias value for 100 results is 0.96, approximately one head for one tail. The results on



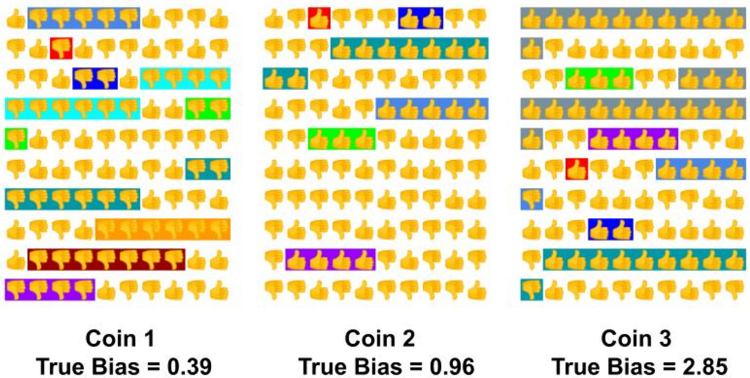
the left can be expected from a coin with bias 0.39, {Picture left top}. There are approximately two tails for each head. The results on the right can be expected if the bias of the coin is 2.85, nearly two heads for each tail. These bias values were calculated using all the 100 results.

In the second picture, {Picture left bottom} a small number of results are highlighted by a dark colored rectangle. If the bias is calculated using the small number of results, the coins on either side appear to be fair and the coin in the center appears to have a bias of 7. To determine correct bias values, the number of results should be as large as possible. Those who challenge modern science argue emotionally, using an extremely small number of observations. The first lesson taught to scientists is to examine the results impartially and not to become very emotional about the conclusions.



If the first few results of a fair coin are heads, would the chance of a tail on the next toss increase? No. To believe that the chance would increase is called gambler's fallacy. {Picture left}

In the results of coin tosses shown earlier, some series of heads and tails have been highlighted by rectangles of different colors. {Picture left below} In the results of the fair coin, one example each of two to nine heads has been highlighted. A tail can follow two heads or three heads. It can also follow nine heads. Alternate heads and tails are however very rare. The highlighted results show that the result of a future toss does not depend on past results. The results on the left are for a coin with a bias 0.39. There are obviously more tails than heads. But even here, a head follows after any number of tails. Series of two to ten tails are all highlighted. But the chance of a head following a head is low. Similarly in the results on the right, of a coin with a bias of 2.85, series with two to ten heads have been highlighted. The bias determines the chance that the result of the next toss being a head or tail. The immediately previous results do not. One has to remember that the coin has no way of storing the results of the earlier tosses. The person tossing the coin may know the results of earlier tosses but



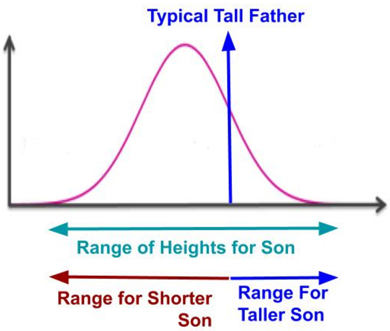
unless there are facilities for cheating, the person cannot change the results obtained by tossing the coin later.

Gamblers use an argument that looks logical. On an average 50% results of a fair coin should be heads. But the previous few tosses are far from the average. So the newer tossed should move the results towards the mean. Newer results moving the mean to the average is called regression to the mean. Knowing the difference between regression to the mean and gambler's fallacy is very important while using scientific knowledge.

The dice with which children pay usually has six sides with numbers 1 to 6 or from 1 to 6 dots. When the dice is thrown, one of the six faces will be on top. There are other types of dice. The ones in ancient India were thin and flat with a dot on one side. So either the side with the dot or the opposite side with no dot could be seen after throwing. Modern dice with a hundred sides numbered from 1 to 100 are also available. {Picture above} In the throw of a fair dice, all options are equally probable.

Regression to the mean is easily explained using the hundred sided dice. If the result of the first throw is 73, the number obtained on the next throw of the dice is more likely to be a smaller number. But if the number obtained on the first throw is a small number like 13 the result of the next throw is more likely to be higher. So at first sight it appears that the numbers are somehow influenced by the past throw. This is called regression to the mean. A large number is followed by a small one, a small one by a large one. But just like a coin, there is no way a dice can store the results. So why is regression to the mean seen in dice throws but not coin tosses?

In coin tosses we are not comparing the head and tail. We are counting the number of heads and tails before comparing them. In the example of the dice, we are comparing the numbers on the dice. That is the difference. When the hundred sided dice is thrown, any number from 1 to 100 is equally likely. But there are only 27 numbers larger than 73 but 72 smaller ones. Obviously a smaller number is more likely. There is a small chance that 73 is followed by 87. But the chance of getting a smaller number in the next throw is even higher. Only 13 numbers are larger than 87.



The relationship between the heights of fathers and sons is a good example of regression to the mean. Sons of a tall father are usually shorter. The probable height of the son could be anywhere in the Gaussian distribution described earlier. But the father is taller than the average of the Gaussian. So the fraction of people in the Gaussian distribution {Picture left} taller than the father will be less. So the chances of a son being shorter are also higher. The comparison with the dice with a hundred sides is quite obvious. But there is a difference. The son could also be tall because he inherited the genes from his father.

A wrong example for regression to the mean can be seen in sports news. The critics will note that the team has unexpectedly lost three of four games and so is more likely to win because of regression to the mean. But the chance of a team winning is similar to the bias of a coin. It is determined by the results of a large number of games. That won't change with the past few results. One can find psychological arguments for either winning or losing the next game. The losses may create despair leading to another loss or a determination to win. But the result has nothing to do with regression to the mean.

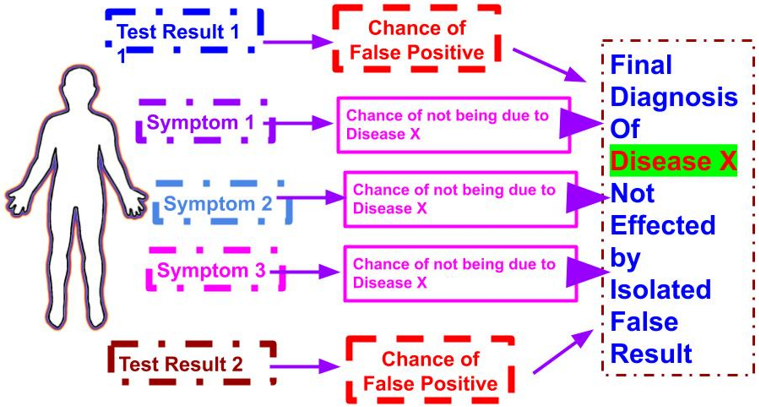
Another important topic in this discussion of the statistics of the results has to be baseline fallacy. As an example consider once again the blood test with a ten percent chance of error. Suppose the test has been employed over a large population. What percentage of people get a wrong result? That percentage will not always be 10%. To understand why this is so, consider the results of the coins with biases 0.96 and 0.39 again. The picture shows the results with a small modification. 10% of the heads are wrongly marked as tails and 10% of the tails are wrongly marked as heads. These have been highlighted. {Picture left} This assumes that there is a 10% chance of recording the tosses wrongly, perhaps because the light was insufficient for proper observation. For both coins, the error is 10%. The bias of the fair coin has not changed. The bias of the other coin has changed from 0.39 to 0.47. Why? In the results of the first coin, 5 heads had became tails and 5 tails became heads. The bias value remains 0.96. But in the case of the second coin there are more tails. So 7 tails became heads and only 3 heads became tails. In practice the error in the results depends not only on the error percentage but also on the bias. This is the cause of baseline fallacy.

Similarly, a fixed percentage of the population will not get wrong diagnosis. If the blood test was for detecting anaemia in poor people, there is a good chance that 50% of the

population are anaemic. So the percentage of wrong diagnosis could still be 10%. But suppose the test is for detecting a rare but dangerous disease. Let us assume further that other symptoms show that 1% of the population has the disease. Now the percentage of wrong diagnosis is far higher. Because of extensive news coverage 100,000 people may get themselves tested. Out of these 1% or 1000 people are genuinely suffering from the disease. But the test will give 10% wrong positives for the 99000 people who do not have the disease. So there will be a fear that 11,000 people are affected and another 100,000 people come forward to be tested. This is the consequence of not recognizing the problems created by baseline fallacy.

Sometimes even scientists fail to take this into account. If an earthquake can be predicted in advance, a lot of people can be saved. Scientists can identify areas where earthquakes are more likely. They can design buildings and bridges that can withstand big earthquakes. But any test that the scientists can devise for detecting an earthquake will have an error. Earthquakes are very rare. So there will be many more false alarms when there is no earthquake than genuine alarms when there is an earthquake. This is the baseline fallacy. This is why predicting earthquakes is impossible in practice. But a group of scientists in Italy strongly advertised a test they developed to predict earthquakes. They ultimately could not predict an earthquake. The local people got angry and used their own publicity as an evidence in a court of law and sent them to jail.

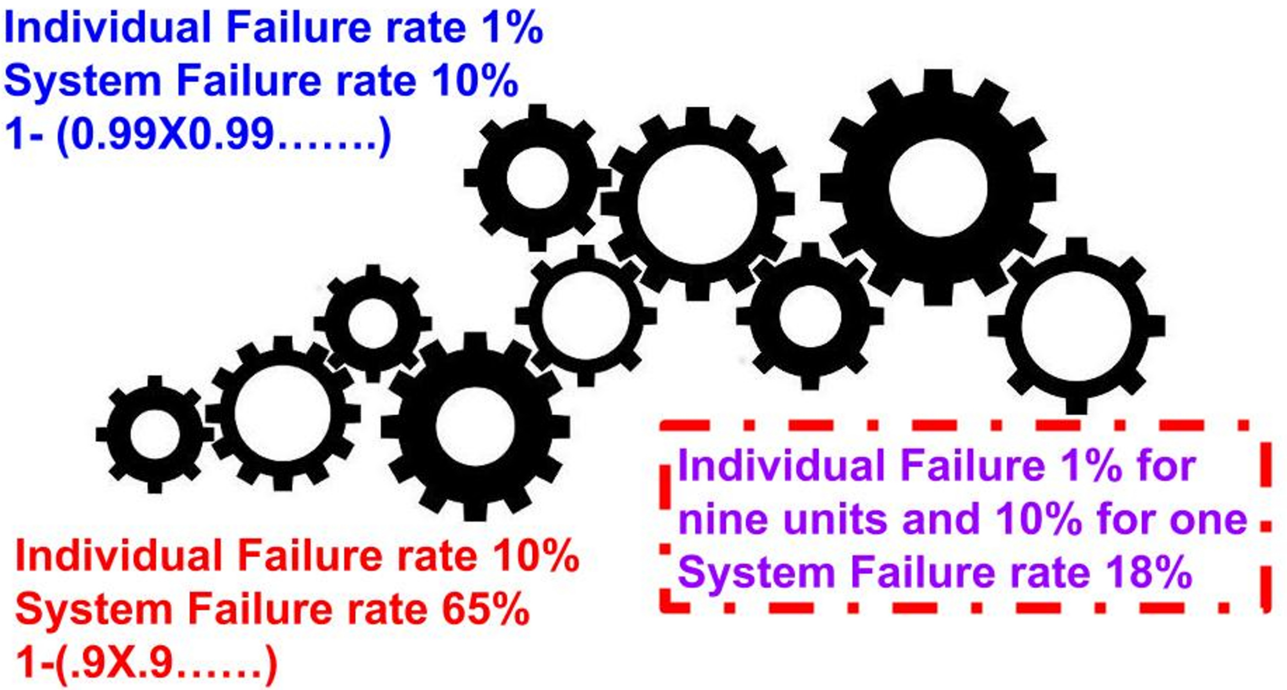
Just as there is an error in every experimental result, there is a chance that any machine would fail. Scientists try to reduce the chance of failure in two ways. A good example for the first method is medical diagnosis. No diagnosis is ever made on the basis of the result of a single test or a single symptom. As discussed earlier, every single result has a chance



of being wrong. So the doctor keeps adding symptoms and tests. The question asked at each step is the same. What is the chance for the patient to have this symptom or for this test to give this positive result without the patient actually suffering from the disease. {Picture left} And finally what is the probability that all these symptoms and all the positive test results could simultaneously be present without the patient suffering from the disease. The doctor is helped by every symptom and test result strengthening the others. Everyone of these must be independent. They join in parallel. When the disease is extremely dangerous, a second opinion is sought. Another parallel addition. This method, called Bayesian statistics was discovered by the 18th century English statistician Bayes.

When the additional expenditure of materials and money is justified, to reduce the chance of failure, similar parallel support is employed in other areas. For example, an air plane is designed to work even if one of the two engines fails. There are two pilots in every plane though one is sufficient.

In simple machines like a wheel, axle or pulley, there are very few parts. As the machine becomes more complex, the number of parts increases. The machine fails to work if any of the parts fails. This has been shown as a chain of toothed wheels. {Picture opposite page}



When the first wheel rotates, so will the last. But only if each and every toothed wheel in the middle also rotates. The last will stop if any of the intermediate wheels fails. This is a serial or chain like design. If the design includes parallel alternatives for each of them, the materials required for making the machine and the energy required to operate it increase making it too costly and unusable. How is the probability of failure of the complete machine related to the probability of failure of the individual units in series? If the chance of failure of

each of the toothed wheels is 1%, is the chance of failure of the machine with ten wheels also 1%? No. The first wheel works 99% of the time. The second works 99% of 99%. So to get the chance of both working, we need to multiply the numbers. Since there are ten such wheels to get the chance of failure of the complete machine, we need to multiply ten times. The chance of failure in this case is about 10%. If just one of the wheels has a 10% chance of failure, rather than 1% like the others, the machine failure rate is 18%. If every wheel has a 10% chance of failure the machine would function only 35% of the time. When the parts are in a chain, they do not reinforce each other. One part cannot perform the job of another that has failed. So in a machine with large number of parts, individual efficiencies have to be extremely high to get a reasonable overall efficiency.

The saying, "A chain is as strong as the weakest link" recognizes this truth. The folk saying that "A rope of grass could easily tie an elephant" recognizes the strength of parallel reinforcement. Science converts such ideas into mathematical relations and numbers. Only then is the design of machines possible.